DOCUMENT RESUME

ED 413 756 FL 024 777

AUTHOR Paskaleva, Elena; Zaharieva, Bojanka
TITLE Tagging a Highly Inflected Language (Non-English, Non-Latin Alphabet, No Morpho Component).
PUB DATE 1995-00-00
NOTE 13p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Alphabets; *Computational Linguistics; Computer Software; *Discourse Analysis; Foreign Countries; *Language Research; *Languages; Linguistic Theory; Program Descriptions; *Programming; *Structural Analysis (Linguistics)
IDENTIFIERS Language Corpora; *SUPERLINGUA

ABSTRACT

This paper describes a computerized system for tagging language corpora that accommodates the special conditions in Bulgarian language research (notably, lack of advanced technology and corresponding user knowledge). SUPERLINGUA has been developed in the DOS environment with consideration for these conditions. Information on product use, users, language versions available or in development, software configuration, functions and linguistic design of the system, and additional operations that derive from the tagging function are described. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 756

Tagging a Highly Inflected Language (Non-English, Non-Latin Alphabet, No Morpho Component)

Elena Paskaleva, Bojanka Zaharieva

Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
1113 Sofia, 25A Acad. G. Bontchev str.
Tel.: +359 2 713 38 41
Fax: +359 2 70 72 73
E-mail: hellen@bgcict.acad.bg

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Norbert Volz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL024777

1. Use of the Product

1.1 General Limitations

The SUPERLINGUA system here presented carries outpresented here implements a special kind of tagging which is far behindlow the power of modern tagging systems. However, as it is pointless to argue that only heavy weaponry is used in a war, in the NLP battle there are also situations where applying all the technology for the processing of linguistic knowledge in corpora is impossible. These situations are of diverse nature, but some of their combinations include:

a) a new, virgin language is processed for which NLP technology has not been developed, or

b) the available technology is inaccessible due to the incompatibility of platforms (hardware or software), and also of institutions (the latter being the hardest to overcome). This partially accounts for the paradox contained in the title: highly inflected language with no morphological component.

1.2 Some Local Limitations

In marketing research, the orientation towards a particular target user is an essential prerequisite for the creators of general software. For NLP products and their application in a highly specialised area, such research is even more important. In countries with a young market economy and serious economic problems, such research is literally a must.

A specific restriction that has to be made is typically Bulgarian and deals with the hardware resources available to the potential users of the system in question - linguists and translators. The marketing research available on computer technology in Bulgaria does not refer to that user-group directly. In a rough outline of the distribution of hardware, banks - currently the most prosperous users - and the Bulgarian linguists, translators, and philological educational institutions obviously occupy the two extremes of the ladder of wealth.

There are many users in Bulgaria who cannot take their pick among different WINDOWS versions (3.1, 3.11 or WINDOWS 95 is unavailable because of their configurations of AT-286 with memory of 2 MB or less) because they do not have powerful enough computers.

In SUPERLINGUA, the chosen hardware configuration and the software platform are crucial for the purpose of outgrowing the boundaries of Research and Development products and discussing, if not an industry, then at least a service to the mass user. This is especially valid for the countries of Eastern Europe with their particular level of computerisation before as well as after the Great Change. The system presented here also has its place in the system of criteria for evaluation of the applicability of a computer product.

1.2.1 Software configuration of SUPERLINGUA and characteristics of the program design

The system claims to really be a working one and to be targeted at the real, i.e., the poor Bulgarian user in question; thus, it has been developed in the DOS environment. With the facilities of that environment, all qualities of the user interface have been developed which are natural functions of every WINDOWS application, including the least extravagant ones.

1.2.2 User and data-type orientation

The system has been designed with the linguist born in mind as the specific user whose hardware deficiency is accompanied by a deficiency of knowledge and skills in computer processing practice. The system is oriented towards the peculiarities of the processed material – large text files and low speed of manual tagging. Options have been developed for that purpose, allowing interruption and resumption of processing with retaining of the intermediate results as well as division of large files into portions. In the process of this arduous work, fool-proof facilities and warning messages have been designed to assist the exhausted or/and artless linguist.

1.2.3 Language orientation

The degree of universality of a certain procedure for extracting linguistic knowledge is determined by the depth of the representation of that knowledge. The only level where procedures can be universally applicable is the text-string level (the normal level of text editors). Even here, the existing discrepancies in character-set representation bring in language-specific

elements. In the DOS environment and the employed DB language (CLIPPER), versions have been developed for Bulgarian, English, and Latin – languages which distribute ASCII space uninterruptedly. A separate version is being developed for Russian texts situated in the Russian coding area, i.e., processed and used in Russia.

1.2.4 Current developments in the software configuration

For the wealthier users, the development of a WINDOWS version of SUPERLINGUA is being prepared with the resources of ClipWin in which many of the specific user-oriented and quasi-WINDOWS features of the system will be included in the very design of the program environment.

2. Functions of the System

The basic function of SUPERLINGUA is the *grammatical tagging* of a text: it carries out the steps in a grammatical analysis of the words in it. These steps as well as all supplementary and accompanying procedures are carried out by the basic modules of the system as follows:
1. Preprocessing module; **2.** Lemmatisation; **3.** Identifying the characteristics of the lemma; **4.** Identifying the characteristics of the word forms of a lemma.

2.1 Basic and Secondary Functions

The basic items and operations of the above mentioned modules allow the application of supplementary procedures: the results, although foreign to tagging itself, have their own significance not only as linguistic operations but as a basis for other important NLP applications. Examples of such an application are:

a) The results of pre-processing of a sufficiently large linguistic database used in an automatic sentence-segmentation and sentence-alignment;
b) The sum of linguistic data obtained from lemmatisation (the transformation of the text to its vocabulary) gives the core of various statistical research and all of the formats of text vocabularies.
c) The developed functions of allowing constant text-support (with editing and revising facilities) also allow for the establishment of all possible types of concordances.

2.2 Principles of Structuring of the System's Linguistic Design

The noble intentions of the SUPERLINGUA's creators are: to provide the linguist with all he wants and is able to receive on the language level within system's reach, in team-work with it. These ideas are materialized in the main principles of the linguistic design, namely:

a) maximum extraction of linguistic information from the current level of presentation and processing;
b) compensation for the lack of knowledge within the system by means of a user-friendly system of menus extracting the user's linguistic knowledge.
c) possibilities for error correction at any stage of the processing;
d) the use of contextual support at all stages and levels of the representation.

The system's linguistic design modeulates the grammatical tagging of the text. This operation which means juxtaposition of the total of text units to the total of those units along with the grammatical information attached to them. There are various degrees in the automatization of tagging, and the two extremes are marked by:

a) completely manual tagging where, for every running word in a sentence, the result of grammatical analysis is entered; therefore, the computer is simply a text-editing tool in that process;
b) a fully-automated tagging which presupposes the availability of a morphological analyser and a disambiguator (which are rule-based or statistical).

Neither can the former operation be completely manual, nor can the latter be fully-automated. The degree of automatization of SUPERLINGUA is the maximum possible automatization of manual tagging where morphological analysis and disambiguation are manual operations assisted and facilitated by the system. Assistance is offered primarily in the structuring of grammatical knowledge.

3. Units and Operations in SUPERLINGUA

In the transition from the original (untouched, virgin) sequence from the input text file (T) to the tagged text (TagT), the following portions of linguistic knowledge are involved, and are considered below according to the levels of knowledge and according to their nature: actual linguistic unit, and information about them.

3.1 Knowledge of the Text Units

In SUPERLINGUA's knowledge of the textual level, the basic unit of representation and processing is the running word (**W**). **W** is a symbol sequence consisting of at least one letter. Other symbol constituents of **W** can be:

- letters: a) belonging to the processed language or to the foreign one;
  b) small or capital;
- figures;
- punctuation marks;
- graphic symbols.

Some criteria based mainly on the formal structure of the concrete **W** distrubute the set of all **W**s into five subsets:

1. regular word forms (**WF**),
2. abbreviations (**Abr**) or
3. capitalized abbreviations (**CapAbr**),
4. proper names (**PropN**), and
5. textual garbage (**Garb**).

The distribution of **W** into the five subsets above is fulfilled by the *preprocessing* module of SUPERLINGUA (Fig. 1). In the preprocessing, the distribution of all **W**s into the sets **WF, PropN, Abr, CapAbr,** and **Garb** is executed in two ways – fully automatically and in a dialogue with a contextual help for the ambiguous cases **W/Abr** and **W/PropN** located on the sentence boundary.

The elements of the first subsetunit (**WF**) provides the connection with the next level of linguistic knowledge – the *morphological* one. This first order distribution is the bridge thrown from the plain reality of to the linguistic representation of the *text* sequence to the ordered reality of *vocabulary* units. In the construction of **WF** vocabulary, the ordering is not a proper linguistic one – it is reduced to a simple alphabetizing and exclusion of repeated elements. The only linguistic knowledge used in a system's dialog is the user's knowledge (supported by the contextual help) in the disambiguation **W/Abr** and **W/PropN** for the boundary cases.

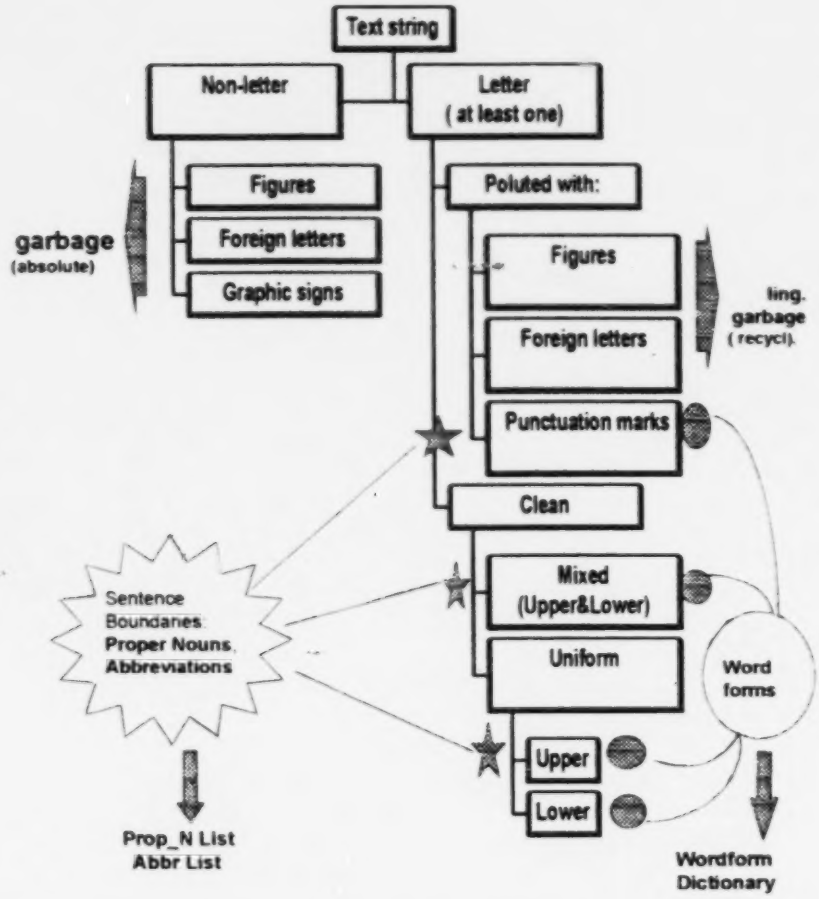

Figure 1. A first order distribution of the text sequence in the preprocessing module

3.2 Knowledge of the Linguistic System

In the system's knowledge of the morphological level of representing the linguistic units, a running word can be a word form (**WF**) or a lexeme (**L**). In the paradigmatic presentation, **WF** and **L** are *variant* and *invariant*, *particular* and *general*. Two corresponding types of grammatical knowledge are formed

during the extraction of linguistic information from the text: the former is *lexical* (**LexInfo**), and the latter is *grammatical* proper (**GramInfo**). **LexInfo** is included in general dictionaries in each *lexical* entry, and the **GramInfo** represents the result of the *grammatical* analysis of each word form obtained through either the Morpho-component or manually (see Figure 2).

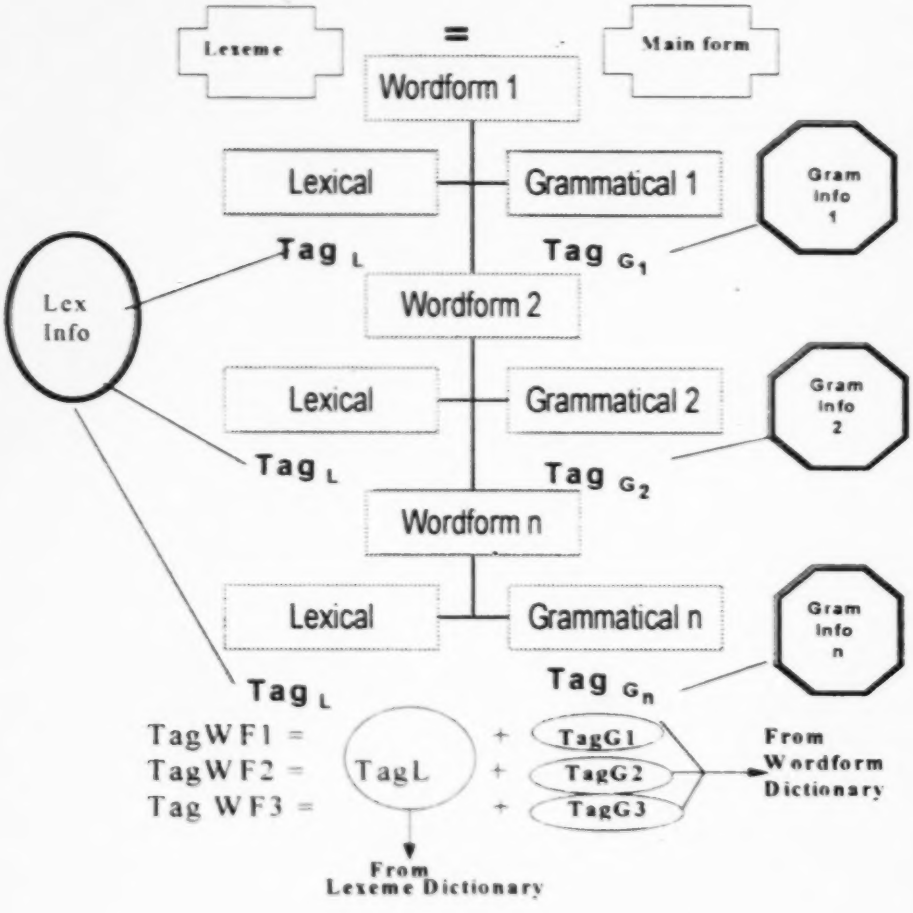

Figure 2. The constituents of tagging information for a highly inflected language

A *tagged* word has to include both types of knowledge (*lexical* and *grammatical*). An optimum organisation of tagging obviously has to provide for the separation of these two types of knowledge because one is obtained for every *class* and the other for every *representative* of the class. The proper distinction of these two types of knowledge (**TagL** and **TagGn**, respectively) is *conditio sine qua non* for every tagset description of a language which is a bit more inflected than English . We can see an unfortunate example of the blending of lexical and grammatical knowledge in the tagset of nine European languages in (1).

The two stages of tagging are – *lemmatisation* (transition from **WF** to **L** with extraction of **LexInfo**) and *grammatical analysis* (assignment of **GramInfo** to every **WF**).

The result of tagging (**TagT**) is a sequence of **TagW**. A **TagW** is a sequence representing **WF** as: **L**, **LexInfo** and **GramInfo**. The main menu of SUPERLINGUA performs the transition **WF–L** with the assignment of **LexInfo** and **GramInfo** to every unit from the **WF** dictionary (the latter is already constructed automatically during the preprocessing) (see Figure 3).

3.3 Optimisation and Acceleration of Manual Tagging

The lack of a morphological component does not automatically lead to the organisation of tagging as processing of *every* word in the textual sequence. The knowledge of the systematic organisation of an inflectional language allows us to accelerate that process.

1. The definition of **LexInfo** as class characteristics saves the repeated assignment of **LexInfo** to *all* class (paradigm) *members*. The more highly inflectional the language is, the greater the economy.
2. Some more time is saved with the help of the minimal morphological knowledge included in the system which contains the constant grammatical features (**GramInfo**) of the word form selected as a basic form in every POS class. The actual tagging of that word form is saved.
3. Also, the transition **WF – L** in the main menu conveniently includes the first part of the next disambiguating procedure – the disambiguation as an operation. Here the *rough disambiguation* is fulfilled, the definition of the lexemes for the homonymous wordforms presented in the processed text (not in the whole language). The context support helps us to correlate the list of the homonymous wordforms (**WFList**) to the list of their lexemes (**LList**) but preceding the tagging. Here, formally coinciding word forms are

attached to different lexemes after contextual help, their place in the text (and the correlation **WF-L**) remains undefined. This is the content of the *real disambiguation* performed on the real text during the next stage (and not on dictionaries of **WF** and **L** processed in the main menu).

Thus, in SUPERLINGUA, tagging of nonambiguous elements is carried out in the dictionary of word forms (**WfD**) and not in the linear text (which saves us *n-1* taggings where n is the number of appearances of the word form). There is no direct manual tagging of the all running words in SUPERLINGUA (see Figure 3).

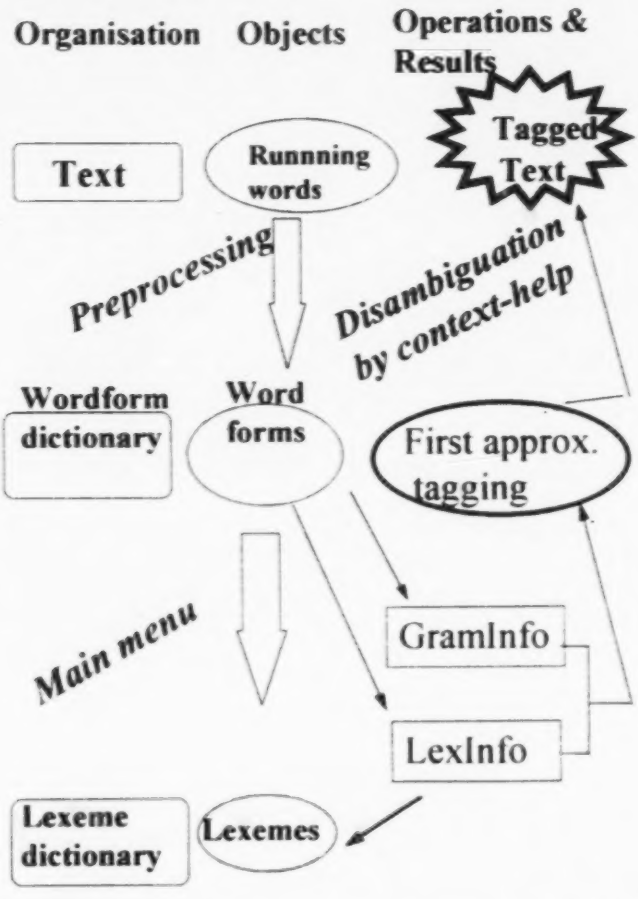

Figure 3. Processed language units, their organisation, and operations in SUPERLINGUA

In this way, the *actual tagging* – processing of the *textual* sequence has to include only:

a) eventual error corrections – in the tagged text conveniently presented for editing, and
b) the second part of the disambiguating procedure – the distribution of the already tagged wordforms in **WfD** in the text (linking the wordforms as *dictionary units* to wordforms as *text units* – appearances).

Apart from the organisation of linguistic knowledge, acceleration of manual tagging can also be achieved through the very organisation of the processing operation. This includes: the division of text files into internal portions of 12KB which are seen by the user only in contextual help; the options allowing interruption of the processing of long files with the options *new* and *next* and their appropriate combination with *save* and *update*.

3.4 Additional Operations

Tagging is the final, deepest function of SUPERLINGUA. Other operations, which utilise the results of pre-processing and the transition **WF-L**, can extract almost all possible combinations of available knowledge and data organisation executed before or along with it. These are common and trivial operations such as *concordances* of various types, the construction production of *frequential* and *reverse vocabularies*, *string search*, standard *set operations* on vocabularies with *statistical data*, etc. Although they do not accelerate tagging directly, they are a useful secondary product of SUPERLINGUA and are manifestations of the above mentioned principle of extraction of all possible information from the available organisation of linguistic data (to give *the linguist all he wants...*)

3.5 Distribution of the System

The designers of the system would like to make their contribution to the fight against terminology abuse and misuse in which the combination of several standard sorting, searching, and string-matching procedures are presented as intelligent NLP achievements (terminology abuse is located on all levels of the linguistic hierarchy (from students' graduate theses to research projects). A well-known way to fight price-speculation is dumping sales. That is why the authors' intent is to distribute the system in the public domain for research.

References

Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicon and Corpora. A Common Proposal and Applications to European Languages. Draft version. EAG-LSG/IR-T4.6/CSG-T3.2. Project EAGLES, October 1994.

Seneca, De ira. 1970. Moral essays, vol. I. The Loeb Classical Library, London p.106